Words or Concepts: the Features of Indexing Units and their Optimal Use in Information Retrieval

Yiming Yang
Christopher G. Chute

Section of Medical Information Resources
Mayo Clinic/Foundation
Rochester, Minnesota 55905 USA

ABSTRACT

Words or Concepts, which are a better choice for indexing the contents of documents? The answer depends on what method is used for retrieval. This paper studies the effects of using canonical concepts versus document words in different retrieval systems with a testing collection of MEDLINE documents. In our tests, for a retrieval system which does not use any human knowledge, using words yielded better retrieval results, while using concepts suffered from a vocabulary difference between canonical expressions of concepts and non-canonical words in queries or documents. For a system which depends on the UMLS synonym set for a mapping from queries or documents to canonical concepts, the retrieval results were slightly better than the case of not using the synonyms, but still worse than the systems using words. For the systems which automatically "learn" empirical connections between words and concepts from examples in the testing collection, the vocabulary problem was effectively solved, and the results of using concepts were competitive or better, compared to those using words.

INTRODUCTION

What is the best way to represent the contents of documents: the documents themselves, or a set of canonical concepts? This is a long-standing controversy in the information retrieval field. While most practical databases use canonical concepts for organizing documents and for indexing the contents, disagreement remains among research-oriented systems. Many of these systems [1] [2] claim that using words in documents leads to best retrieval performance, and others [3] believe in the use of concepts for more accurate indexing.

In recent years, the development of the Metathesaurus of the Unified Medical Language System (UMLS) of the National Library of Medicine [4] has been drawing research attention to the potential and optimal use of canonical concepts. SAPHIRE [5], for example, is a system designed for using the UMLS thesaurus for automatic document indexing and retrieval based on canonical concepts. SAPHIRE was recently compared [6] with alternative methods which do not use canonical concepts but rather document words as indexing units. In this evaluation of SAPHIRE, Hersh observed that "The consistently best methods are those that use indexing based on the words that occur in the available text of each document. Methods used to map text into concepts from a controlled vocabulary showed no advantage over the word-based methods."

We found Hersh's test interesting. Our question is, what is the major reason behind the poorer performance of using concepts? The potential reasons are:

(1) the UMLS "main concepts", i.e. the MeSH subject categories (Medical Subject Headings, or MeSH) [7], are not rich enough or precise enough for identifying the contents of the documents; or

(2) the UMLS synonyms and lexical variations are not rich enough to cover the vocabulary of the queries and the documents.

Finding the true reason(s) is important. If the first answer is the reason for the poor performance, then the only hope for improvement is to find a better set of concepts to replace the UMLS collection, or to use words instead of concepts. If the first answer is not true and the second answer is true, then either enriching the UMLS synonym set or using an example-based learning approach for an empirical mapping between different vocabularies would lead to an improvement.

In this paper, we will verify these hypothetical reasons through tests on a MEDLINE document collection. Our study consists of two parts:

Part 1. Testing the retrieval effectiveness of using *concepts* versus *words* under the condition that the mapping among different vocabularies (controlled and non-controlled) is solved. That is, we separate the effect of semantic representation from the effect of mapping among surface expressions.

0195-4210/92/$5.00

We choose the Linear Least Squares Fit (LLSF) mapping method for such a purpose. The LLSF mapping is an example-based approach [8] [9] [10] which automatically learns from a training set of relevant queries and documents the word-to-concept connections when a document is represented as concepts, and word-to-word connections when a document is represented by its own text. These connections have the functionality of terminology thesauri without requiring human effort in developing synonyms. Having the vocabulary difference between queries and documents solved by the LLSF mapping, we can observe the effect of the choice of semantic units, *concepts* or *words*, by comparing the difference in retrieval performance, if any.

Part 2. Evaluating the effectiveness of typical retrieval methods in solving the problem of vocabulary differences among queries, documents and/or canonical concepts. The tests include a word-based matching method which does not use any human knowledge, a concept-based matching method which employs the UMLS synonyms, and two example-based approaches which learn from relevant queries and documents. We choose the SMART [1] system without the relevance feedback part for testing word-based matching. We cite the published results of SAPHIRE for the study of using the UMLS synonyms. For the example-based mapping, we compare the LLSF with SMART using relevance information. Observing the results, we can have a quantitative analysis on the vocabulary problem and the effectiveness of the solutions.

THE TESTS OF THE LLSF MAPPING

The Testing Data

The MEDLINE document collection chosen for our test was the largest of the testing sets used in the evaluation of SAPHIRE by Hersh [6]. This collection was originally designed by Haynes and McKibbon for an evaluation of MEDLINE [11]. The original set consists of 78 queries and 3,403 citations. A citation is a data entry in MEDLINE, each containing a title and/or an abstract, and a set of subject categories (MeSH terms) assigned by human experts from the National Library of Medicine. Adapting our terminology to this, we call the title and the abstract together a document, and the subject categories the concepts. Hersh reduced the original testing set by eliminating documents in which an abstract was not present and queries which did not have relevant documents in the reduced document collection. The resulting set has 75 queries (the Novice queries) and 2,344 documents. We will refer to the reduced testing set as the Shared Testing, because we use this set for our test.

The LLSF approach requires a training set, that is, a set of matched queries and documents. The 2,344 documents in the Shared Testing set contain 991 documents (42%) which are relevant to the query set and 1,353 documents (58%) which are irrelevant to any of these queries. We split the data into a training set and a testing set. We sorted the relevant query/document pairs (1,074) by document, and took the documents in the odd pairs for training, and the other documents for testing. The resulting training set contains 71 queries and 524 relevant documents, and the testing set contains 68 queries and 1,820 documents. Only 22% of the testing documents are relevant to the testing queries and 78% are irrelevant. Eighty-eight percent of the testing queries are contained in the training set, but there is no overlap among the training documents and the testing documents. We call the training set "Disjoint Training" and the testing set "Disjoint Testing".

We split the data this way so that most of the testing queries can use the term-to-concept connections obtained from the training set, but none of the known answers are included in the testing set. The test is to verify how much the LLSF mapping can capture unknown matches (documents) after training from the knowns. An alternative choice is to use the Shared Testing set instead of the Disjoint Testing set for the evaluation. The result of the Shared Testing set would be better because the retrieval algorithm would favor the documents contained in the training set (roughly 50% of the total relevant documents). However, it would not make much sense to count the known answers as a part of the retrieval achievement, and it would be unfair if we compare such a result with other approaches in which none of the answers are known before the retrieval.

The Results

For evaluating retrieval effectiveness, we use the conventional measures, recall and precision.

Definition. The recall and precision of a retrieval with respect to query q are

$$\text{recall } (q) = \frac{\text{number of documents retrieved and relevant to } q}{\text{total number of documents relevant to } q}$$

$$\text{precision } (q) = \frac{\text{number of documents retrieved and relevant to } q}{\text{total number of documents retrieved}}$$

For a set of queries, we compute the recall and precision for each query and then average them: for recall threshold at 10%, 20%, 30% ... 100%, retrieve as many documents as needed for each query, and average the precisions of the points where the threshold is achieved.

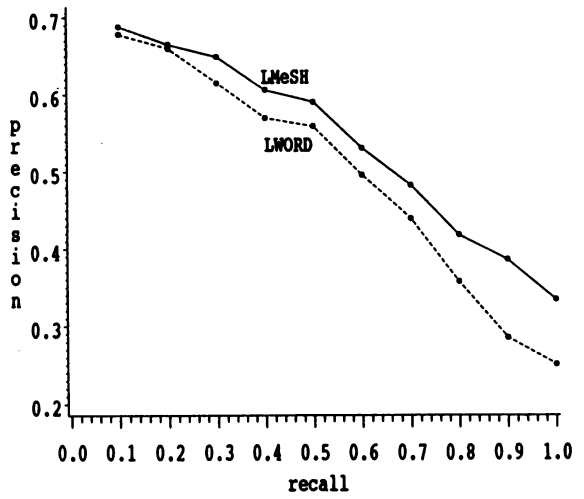

Figure 1. Using concepts versus words in the LLSF mapping

We represented the documents by MeSH concepts and by document words and tested the LLSF mapping on each of them. The two tests are named LMeSH and LWORD. Figure 1 shows the recall-precision curves of LMeSH and LWORD on the Disjoint Testing set. The performance of LMeSH was slightly better than LWORD. We do not consider this a significant difference, compared to the differences among the LLSF approach and the alternative retrieval methods as we will see in the later discussions. Using concepts as opposed to words did not make any significant difference because the mapping among different vocabularies was well solved in the LLSF approach. The empirical mapping function is equivalent to word-to-concept connections in the case of using concepts, and equivalent to word-to-word connections in the case of using words.

On the other hand, what made the performance of LMeSH slightly better? A potential advantage of using concepts for document indexing is the elimination of non-informative or "noise" words, compared to using all the words that occur in the documents. A potential tradeoff is that the concepts might be too general and not reflect the contents of particular documents as precisely as using the original words. Nevertheless, Figure 1 shows that the overall performance of concepts was better. Although the MeSH concepts were not developed for this particular testing set, they have better performance than the original document words; the MeSH concepts are rich enough and precise enough for representing the contents of the documents.

THE TESTS ON DIFFERENT METHODS

We have explored the use of concepts and words in the situation where the mapping among different vocabularies was solved by an example-based approach. Now we study the situations where alternative retrieval methods are used. The focus is on how much the vocabulary problem affects the retrieval effectiveness of different approaches. This comparison includes the following tests:

SMART: a test of using document words as indexing units in a word-based approach to retrieval. For this test, we ran the SMART system developed by Salton's group [1] and recognized as one of the most representative retrieval systems. Its basic approach is a word-based matching among queries and documents, with use of statistical word weights; no human knowledge is required in this approach.

SMeSH: a test of using MeSH words as indexing units in a word-based approach to retrieval. That is, the documents are represented by the words that occur in the corresponding MeSH concepts. We ran SMART for this test, and named it SMeSH.

SAPHIRE: a system using the UMLS main concepts for indexing and employing the UMLS synonyms (78,244) for mapping queries and documents to these concepts. SAPHIRE uses a phrase-based matching algorithm. The result is from the published data by Hersh [6].

LMeSH: a test of using MeSH concepts as indexing units in an example-based approach, as described in the previous section.

SMART+: a test of using document words as indexing units in an example-based approach to retrieval. We modified the relevance feedback scheme of SMART into a version that does not require user feedback for identifying the relevant documents for each query. We used a training set, the same one we used for the LLSF mapping, where the relevance among queries and documents is given. We expanded each testing query by adding the words of the relevant documents (if any) in the training set, and then used the expanded queries in the retrieval. We did not apply the relevance feedback part of SMART because that would use different relevant documents for query expansion, and this would make the comparison with the LLSF method difficult. Since our focus is on the effectiveness of relevance information and not on the user interac-

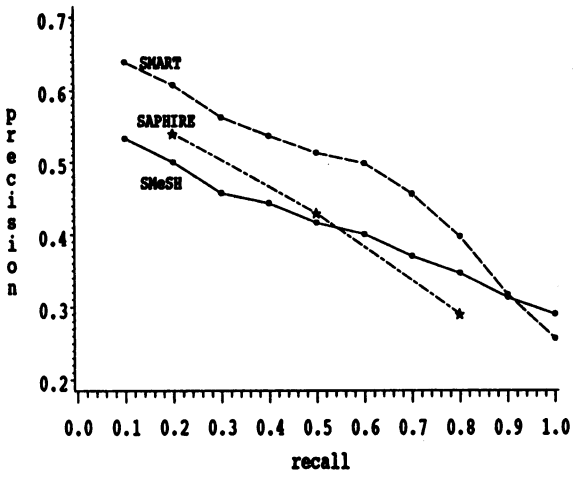

Figure 2. Different retrieval methods on the Shared Testing set

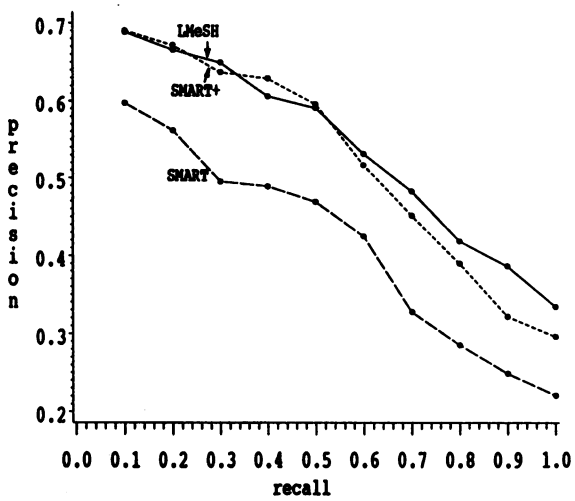

Figure 3. Different retrieval methods on the Disjoint Testing set

tion part of SMART, this modification would not be inappropriate.

Figure 2 shows the results of the methods on the Shared Testing set, including SMART, SAPHIRE and SMeSH. In the published results of SAPHIRE, only the precisions at recall of 20%, 50% and 80% were available, so we plotted those three points. Figure 3 shows the results of the methods on the Disjoint Testing set, including SMART, SMART+ and LMeSH. Since two different testing sets were used, we cannot make a direct comparison of all the results. So we ran the SMART system on both sets for an indirect comparison. Observing that the curve of SMART on the Disjoint Testing set is worse than its curve on the Shared Testing set, we can say that Disjoint Testing is more difficult than Shared Testing. This can also be observed from the percentage of irrelevant documents, that is, 78% in the Disjoint Testing set versus 58% in the Shared Testing set. In the comparison of the results, the relative difference is more important than the absolute values.

Combining Figures 2 and 3, we can make some observations about the effect of the vocabulary problem in the different approaches.

SMeSH reflects a typical situation of retrieval in practical databases where documents are indexed by subject categories (concepts) or keywords, and a word-based search is used for the retrieval. A vocabulary difference between queries and concepts is a crucial problem in such an approach.

SAPHIRE, employing a very large collection of the UMLS synonyms (78,244), had a performance between SMeSH and SMART only on the high precision end, and poor performance elsewhere. This shows that the UMLS vocabulary at the concepts level has a poor fit to this particular testing set. In other words, for systems that depend on general-purpose terminology thesauri, mapping from arbitrary queries and documents to a controlled vocabulary remains a bottleneck problem of using concepts.

SMART, with better performance over both SMeSH and SAPHIRE on one hand, is significantly worse than SMART+ and LMeSH. This shows the limit of the word-based retrieval method. A common weakness of word-based approaches is that they ignore the information within non-shared words in queries and documents, and therefore have limited success when different words happen to be used for relevant concepts by database users and article authors. In other words, vocabulary difference is not only a problem in using concepts; it is also a problem in using words. Solving this problem is beyond the power of word-based matching.

LMeSH and SMART+ had very similar performance in the retrieval and both have a significant improvement over SMART. The mapping among different vocabularies was effectively solved by the word-to-concept connections in the LLSF mapping, and by expanding queries to broader sets of words in the SMART+ approach. There is a common feature between these two methods in that they both use empirical connections between queries and documents, and these connections come from human knowledge about relevance. On the other hand, the two methods also have significant differences, such as the ability to handle ambiguities and to preserve context sensitivity of the mapping; these features are studied in a separate paper [10]. The differences, however, do not have much effect to this test.

DISCUSSION

We have studied the dual roles of indexing units (concepts and words) in document retrieval. Indexing units serve as semantic units for representing the contents of documents, on the one hand, and as lexical units for mapping among surface expressions, on the other hand. The semantic representation requires the indexing units to be complete, precise and non-ambiguous in meaning, while the lexical mapping requires a broad vocabulary coverage. Our tests show that the MeSH canonical concepts are competitive or better as semantic units, compared to document words. However, as lexical units, the effects of using concepts as opposed to words depend on what method is used for retrieval.

For word-based retrieval which does not use any human knowledge, document words had better performance than the controlled vocabulary of canonical concepts, because of the broader vocabulary of documents. However, a major weakness of word-based methods is that the information within the non-shared words in queries and documents is ignored. Given that people use a variety of expressions for particular concepts, it is impossible for database users to figure out all the words likely to be used in the relevant documents. A satisfactory retrieval therefore is unlikely if the vocabulary mapping remains unsolved.

The example-based approaches effectively solved the vocabulary mapping problem by using human assigned relevance. However, the other approach to vocabulary mapping, the method of using a terminology thesaurus, had poor performance. Given that the use of human knowledge is essential in both of the example-based and the thesaurus-based methods, the different performance raises a basic question, that is, what kind of human knowledge is necessary for an effective retrieval?

The example-based methods use a training set from the same application as the testing set, so the empirical connections between different vocabularies are self-restricted to be domain specific, application specific and user group specific, and usually have a better fit to the application than a general-purpose thesaurus. This suggests a potential improvement in thesaurus development, that is, a marriage of the example-based learning algorithm and human refinement. We are optimistic about a data-driven development of application specific components of terminology thesauri and a selective use of these components with respect to applications.

Acknowledgement

We would like to thank William R. Hersh for generously providing the testing data. This work is supported in part by NIH support grants LM-07041, LM-05416, and AR30582.

References

1. Salton G. Development in Automatic Text Retrieval. *Science* 1991:253:974-980.

2. Salton G, Buckley C. Global text matching for information retrieval. *Science* 1991:253:1012–1015.

3. Evans DA, Hersh WR, Monarch IA, Lefferts RG, Handerson SK. Automatic indexing of abstracts via natural-language processing using a simple thesaurus. *Medical Decision Making* 1991;11/4 Suppl;108-115.

4. Lindberg D, Humphreys B. The UMLS knowledge sources: tools for building better user interfaces. *Proc 14th Ann Symp Comp Applic Med Care (SCAMC 90)* 1990;14:121–125.

5. Hersh WR, Haynes RB. Evaluation of SAPHIRE: an automated approach to indexing and retrieving medical literature. *SCAMC 91,* 1991;15:808–812.

6. Hersh WR, Hickam DH, Leone TJ. Words, concepts, or both: optimal indexing units for automated information retrieval. *SCAMC 92,* 1992;16:644–648.

7. *Medical Subject Headings (MeSH).* Bethesda, MD: National Library of Medicine, 1993.

8. Yang Y, Chute CG. A Linear Least Squares Fit mapping method for information retrieval from natural language texts. *Proc 14th International Conference on Computational Linguistics* 1992:447–453.

9. Yang Y, Chute CG. An application of least squares fit mapping to clinical classification. *SCAMC 92,* 1992; 16:460–464.

10. Yang Y, Chute CG. An Application of Least Squares Fit Mapping To Text Information Retrieval. *Proc 16th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval* 1993; 281-90.

11. Haynes R, McKibbon K, Walker C, Ryan N, Fitzgerald D, Ramsden M. Online access to MEDLINE in clinical settings. *Ann. Int. Med.* 1990;112:78–84.